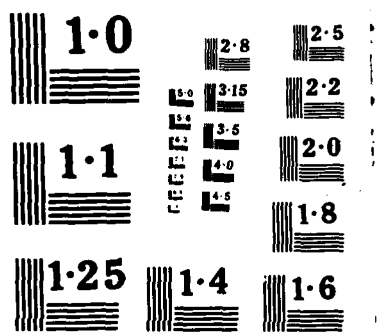
AD-A187 531 WHEN DECISION AIDS FAIL(U) HARRY G ARMSTRONG AEROSPACE MEDICAL RESEARCH LAB WRIGHT-PATTERSON AFB OH E O EIMER MAR 87 AAMRL-TR-87-035 1/1

UNCLASSIFIED F/G 23/2 NL


1·0
2·8
2·5
3·15
2·2
3·5
2·0
1·1
4·0
4·5
1·8
1·25
1·4
1·6

AAMRL-TR-87-035

WHEN DECISION AIDS FAIL (U)

ERHARD O. EIMER

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY

MARCH 1987

FINAL REPORT FOR PERIOD JUNE 1985 - JUNE 1986

Approved for public release; distribution is unlimited.

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY
HUMAN SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573

NOTICES

TECHNICAL REVIEW AND APPROVAL

AAMRL-TR-87-035

The voluntary informed consent of the subjects used in this research was obtained as required by Air Force Regulation 169-3.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

CHARLES BATES, JR.
Director, Human Engineering Division
Armstrong Aerospace Medical Research Laboratory

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (*If applicable*)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS (*City, State, and ZIP Code*)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.
	62202F	7184	27	02

11. TITLE (*Include Security Classification*)

When Decision Aids Fail (U)

12. PERSONAL AUTHOR(S)

Eimer, Erhard O.*

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Final	FROM 850615 TO 860615	1987 March	32

16. SUPPLEMENTARY NOTATION

*On Sabbatical Leave from Wittenberg University

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	
05	05		C^3-system, Communication, Decision aids
05	10		Judgment, Uncertainty

SUMMARY

This report describes an effort to ascertain how a decision aiding device (DAD) affects human performance in a C^3-task that requires judgment about uncertain events. By means of a signal detection analysis, rating data were transformed into measures of task performance, of reliance on the DAD, and of observer efficiency. It was found that performance was a direct function of DAD validity. The manipulation whether the DAD was machine or human co-observer information did not affect performance. Performance with a low-validity DAD was inferior to performance without a DAD. Under all conditions, subjects performed at a level inferior to that which could have been achieved with a maximixing strategy. When DAD validity was reduced to chance level (simulating a breakdown of the DAD), observer performance was reduced below the level achieved by a control group without a DAD. Performance with an invalid DAD was equally poor regardless of previous DAD validity. The considerable variation between observers after DAD breakdown was not attributable to previous DAD characteristics. Instead, performance after DAD breakdown was found to be highly correlated with the efficiency which the subjects showed in their previous use of a valid DAD. Efficiency is not the same as performance. Rather, it can be conceptualized as the ability to use valid information and to ignore invalid information.

PREFACE

The research underlying this report was conducted while the author was a member of the Harry G. Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, during a sabbatical leave from Wittenberg University. The support of Charles Bates, Walter Summers, and Kevin Holloran is gratefully acknowledged. The informed and substantial contributions of Sharon L. Ward and of R. J. Poturalski were essential to the success of this research. Special mention is made of the creative, skilled, and meticulous efforts of Danny Bridges, who wrote the computer software for the experimental task.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION 1

THEORETICAL CONCERNS IN THE USE OF DECISION AIDS

Decision aids abound in our society, and in particular in technically complex command, control, and communications (C^3) systems. Any device that transforms, summarizes, or displays information in some automatic manner may be thought of as a decision aid. Road maps, clocks, odometers, traffic signs, yard sticks, price lists, calculators, and weather forecasts are among the decision aids used regularly. A look at the clock aids in making the decision whether there is enough time for a second cup of coffee, a road map supports the process of deciding where to turn off a highway, and the weather forecast will contribute to the decision whether to pack an umbrella. While some decision aids are transparent and easily understood, others are not. Similarly, there is considerable variation in the degree of fallibility that is peculiar to decision aids. One is more inclined to trust in the accuracy of a traffic sign than of a weather forecast.

C^3-systems are characterized by the fact that they process substantial amounts of interacting information, usually with the objective to make and to support decisions. The number of decision aids in such systems is considerable. These aids often are quite complex, and their function may not be well understood by the operators who use them. In addition, the accuracy of the decision aids may be less than desirable, as well as unknown. In fact, the accuracy may change, possibly without warning.

CHARACTERISTICS OF INFORMATION USE

The effect of decision aids on the outcome of the decision process is not understood very clearly. In particular, it is not well known whether human operators use decision aids to greatest advantage, or whether they may actually be misled by a decision aid. Direct evidence supporting such concerns is less than abundant, but insights gained from a related field suggest that there may be reason for systematic study.

In 1956, Festinger, Riecken, and Schachter published their insightful social psychological study <u>When prophesy fails</u>. The book was received with more emotion than is

common for scientific studies, possibly because it questioned some of the most basic assumptions and foundations of our society. In particular, the authors noted that human behavior does not adjust to disconfirmation of predictions in the manner of a rational decision maker. Rather, the human is likely to expand the original prediction to encompass the manifested disconfirmation, even if this expansion amounts to blatant internal and external inconsistencies. Neither evidence nor argument will abolish a prophesy. This behavior pattern is not restricted to a particular subgroup of humans.

There is but a limited similarity between prophesies such as those studied by Festinger and his colleagues (1956) and the output of a decision aid. Even if the algorithm underlying a decision aid is not obvious to the user, it is based on some rational model rather than on inspiration. Nonetheless, it is not self-evident that humans would be maximally rational in the use of a decision aid when there is abundant evidence that they are not maximally rational in a great variety of other contexts. The peculiar adjustment to disconfirmation of a prophesy is merely one of many examples of suboptimal behavior. People underestimate high probabilities and overestimate low ones; they attribute causality where there is none; they are more confident than they should be on the basis of data available to them; they gamble when it would be in their best interest not to gamble; they believe they have control over chance events; and they exhibit biases of this variety even after extensive efforts to inform them of these patterns and thus to debias them (Kahneman, Slovic, and Tversky, 1982). In sum, there is reason to subject to careful analysis the manner in which human operators make use of decision aids.

DECISION AIDS IN C^3-SYSTEMS

In a C^3-system, observers commonly have to make decisions in a context in which (a) incomplete information is available, (b) decision aids in the form of computer predictions, co-observers with different, relevant information, or both are present, and (c) information sources which are normally valid may become invalid, either through system defects or through faulty information. Observers who become accustomed to highly valid decision aids are likely to rely heavily on these aids, and it is likely that system failure, due to the unjustified reliance of an observer on a suddenly invalid decision aid, is most severe if the decision aid previously was highly valid. If

the decision aid is seen as a machine rather than as a human co-observer, the problem may be more serious, assuming that human observers consider co-observers more fallible than a computer. It appears likely that the observer will re-evaluate the merits of trusting a failing co-observer more readily than those of relying on a machine.

Decision aiding devices such as ADDAM (Adaptive Dynamic Decision Aiding Mechanism), developed by Freedy and collaborators (i. e., Freedy, Davis, Steeb, Samet, and Gardiner, 1976), have shown substantial effectiveness in improving the decision quality of human observers. These sophisticated decision aids not only maximize effective use of available data by providing a stable memory and an optimal information aggregation mechanism to the observer, but they also observe the decision behavior of the decision maker and derive from it a set of estimated utilities for different alternative decision outcomes. Aids of this sort assist in applying the decision maker's own decision strategy more consistently. A major strength and limitation alike that is inherent to such an "intelligent" system is that it must learn, over a number of trials. Such learning is beneficial in that it allows the inclusion of situational features that might not have been incorporated in a mechanism based on a normative model. It is, however, also problematic in that it can be developed only if there is opportunity for repeated learning trials. Many events for which a decision aid might be most crucial are not repeatable. Simulations, exercises, and maneuvers do not serve as satisfactory substitutes for "real" learning trials.

A problematic feature inherent in the use of even the most sophisticated decision aids is that the human operator may become dependent on them, thereby losing effectiveness in those circumstances in which observer effectiveness is most particularly required, namely when the decision aid breaks down or uses incorrect information. It is possible that this problem, in fact, is most serious in those cases in which the decision aid is most sophisticated and has previously been most reliable.

CO-OBSERVERS AS DECISION AIDS

As was suggested previously, it is possible to consider even a human co-observer as a decision aid. In fact, consultation with fellow workers is possibly the most commonly sought aid a person is likely to seek before

arriving at a decision. It is of interest, of course, whether human decision aids have an effect on the observer that is different from the effect of a non-human decision aid that would be comparable in all other regards.

In signal detection terminology (Swets, 1964, Green and Swets, 1966), it is possible that information regarding a co-observer's decisions results in a criterion shift rather than in improved team performance. Under such circumstances a group decision making approach might work against the achievement of maximum system efficiency. There is indeed evidence that group (or peer) pressure can affect decisions profoundly (Asch, 1956). There also is evidence that group decisions tend to be more polarized than decisions made by individuals (Pruitt, 1971), frequently in a risky direction (Wallach and Kogan, 1965), even when there is no rational basis for the decision (Greenberg, 1979). Conversely, there is some evidence (Bateson, 1966) that co-observer decisions may serve an informational function, primarily by focusing the observer's attention on the relevant features of the task. Maier (1970) found enhanced group performance, but specifically in creative and problem solving tasks. Hill (1982) showed "that group performance was generally qualitatively and quantitatively superior to the performance of the average individual. Group performance, however, was often inferior to that of the best individual in a statistical aggregate and often inferior to the potential suggested in a statistical pooling model" (p. 535). Fischer, Edwards, and Kelly (1978) found evidence supporting the view that group decisions are generally inferior and show a conservatism effect when compared with statistical averages of decisions made by isolated individuals. The analytic work of Shiflett (1979) suggests that the relatively poor decision making performance of groups may be the product of the increased task complexity resulting from group structure. These and other sources of evidence would suggest degradation of performance for cases in which an observer has information regarding the decisions of a usually accurate co-observer.

STUDY PLAN

In view of the issues discussed above, the present research assessed how human decision making is affected by three variables: (a) the nominal validity of a decision aiding device (DAD), (b) the nature of the DAD (human co-observer information or automated machine information), and (c) reduction in validity of a DAD. Subjects were given

discrete judgment tasks, each of them with dynamic time constraints, in which they had the support of a human or machine DAD of predetermined validity. After a period of experimental exposure, the validity of the DAD was reduced to chance level. By means of a method based on signal detection theory the decision behavior of the subjects was analyzed for three features, namely decision making performance, reliance on the decision aid, and efficiency of decision aid use. It was expected that observers rely most on highly valid decision aids, that they rely more on machine than on human DADs, and that performance is degraded most severely when a previously highly valid DAD becomes invalid.

SECTION 2

METHOD

To accommodate the requirements developed in the previous section, pairs of subjects were exposed to the experimental procedure in separate rooms, with comparable equipment. As a function of the experimental condition to which subjects were assigned, they were operating either with or without the support of a DAD. If they had DAD support, they were instructed either that the DAD was their team member, or that it was a computer algorithm. On each of a series of trials, subjects had to predict the outcome of a tracking display.

EXPERIMENTAL DESIGN

The research plan called for a mixed three-factor design of treatments (factor A) by treatments (factor B) by trial blocks (factor C) by subjects, factor C being a repeated measures factor (See Table 1). Factor A had three levels of initial DAD validity (to be discussed below). Factor B had three levels regarding the nature of the DAD (machine, human co-observer, or none, i. e., no DAD). Since DAD validity cannot be manipulated as an independent variable when no DAD is present, the design can also be regarded as a factorial design with a single control group (Winer, 1971). This control condition provided an empirical baseline. (A second, theoretical baseline was provided by the a priori validity of the DAD).

Factor C consisted of eight blocks of 40 trials each. These eight blocks were broken down into two stages (I and II). In stage I, consisting of the first five trial blocks, DAD validity was, depending on the level of factor A, moderately to very high. During the three trial blocks of stage II, DAD validity was reduced to zero.

SUBJECTS

Subjects were Air Force military and civilian volunteers and paid volunteers recruited from a large state

TABLE 1. Experimental Design, Factors A and B

Factor A: Validity of DAD	Factor B: DAD Type B_1: Machine	B_2: Human	B_3: None
A_1: r = 0.40 (low)	$n_{1,1} = 4$	$n_{2,1} = 4$	
A_2: r = 0.60 (medium)	$n_{1,2} = 4$	$n_{2,2} = 4$	$n_{3,*} = 12$
A_3: r = 0.80 (high)	$n_{1,3} = 4$	$n_{2,3} = 4$	

university. The subjects were between 18 and 30 years of age. Assignment to conditions was based on the order in which subjects signed up for the study and on scheduling considerations. Care was taken that subjects working together had at most a passing acquaintance with each other before the study.

EQUIPMENT

A PDP 11/34 computer with VS11 graphics system was used for experimental control and data collection. Responses crucial to the problem solving task were manual depressions of pushbuttons on two identical pushbutton response panels, one for each of the two subjects working at any time. Each response panel contained six microswitch push buttons in side-by-side arrangement. The problem task was presented to the subjects on two color monitors, one for each subject. The subjects were located in separate rooms, and they were monitored by color video cameras. Subjects could not hear each other. Depending on the experimental condition, one of the subjects could see the other one (the first subject's human DAD) on a video monitor, without seeing the DAD operating the response panel.

THE TASK

Each trial lasted 10 sec., with an intertrial interval of 2 sec. On each trial subjects saw a visual showing the paths of two vehicles traveling on a potential collision course, at constant velocity and along a straight line. At the onset of a trial both vehicles (depicted as small moving circles that left "tracks") started from the two lower corners of a square field toward a point of intersection. The angle of intersection between their paths was between 53 degrees and 90 degrees (See Figure 1a). A collision could occur between 5 and 10 sec. after trial onset. Collisions occurred, under all experimental arrangements, with a probability $p = 0.50$.

While subjects could see the elongating track of one of the two vehicles throughout the trial, the other vehicle track was developed only for the first two seconds. The vehicle and its remaining course track then became invisible (See Figure 1b). Simultaneously with trial onset, a signal in the lower center of the display indicated to the subject whether a DAD predicted a collision (except, of course, in the condition in which no DAD was offered).

Subjects were instructed that the DAD had similar information available, except that this information was given to the DAD before the subject received it. As soon as the DAD had made a decision, the trial began for the subject, so that the DAD and the subject made decisions on the same task in tandem form, the subject knowing what the previous DAD decision was. Subjects were instructed to make a decision as quickly as possible after trial onset, before a collision could occur. If the subject had not made a decision response within the first four seconds, a buzzer sounded to convey to the subject that a response was overdue. The subjects made the decision response by pushing one of six microswitch push buttons denoting the spectrum from "I am sure there will be a collision" to "I am sure there will not be a collision". As soon as the response was made, the obscured vehicle and its track were restored (Figure 1c), and both vehicles continued either to their collision or to the upper field border (Figure 1d).

The task was fairly difficult since even those trials that did not result in a collision brought the vehicles to a near miss. On these trials, they would cross the intersection of their paths with a time difference of one second.

During the five trial blocks of stage I, the validity of the DAD had values of $r = 0.40$, 0.60, and 0.80 for the

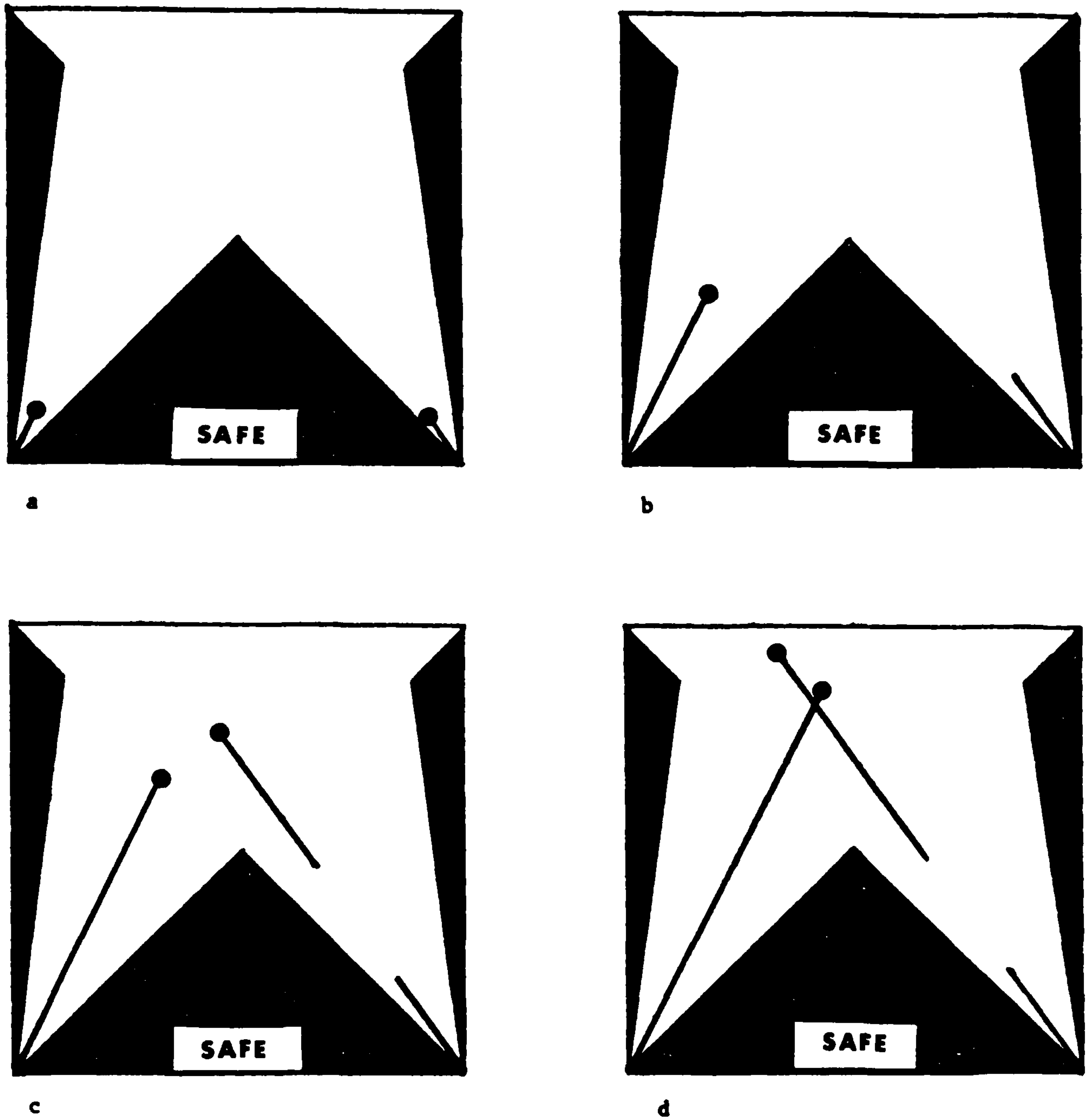


FIGURE 1. Display seen by the subject: (a) 1 sec. after trial onset, (b) 2 sec. later (note that the right-hand vehicle has disappeared and the track has stopped), (c) after the subject has made a response, and (d) after the vehicle paths have crossed without the vehicles colliding.

three levels of factor A. The frequencies of agreement and disagreement between the DAD and the trial outcome that correspond to these coefficients are summarized in Table 2, Treatments A_1, A_2, and A_3. In each trial block, there were 20 collision trials and 20 non-collision trials, in random order. Also within each block, there were 20 trials on which the DAD predicted a collision and 20 trials on which it predicted safety.

TABLE 2. Frequencies of Trial Conditions Per Block as a Function of DAD Validity (r)

Treatment (Level of factor A)	r (Validity of DAD)	Prediction of DAD	Trial Outcome	
			Collision	Safety
A_1	0.40	Collision	14	6
		Safety	6	14
A_2	0.60	Collision	16	4
		Safety	4	16
A_3	0.80	Collision	18	2
		Safety	2	18
X	0.00	Collision	10	10
		Safety	10	10

During the last three trial blocks (stage II), DAD validity was reduced to $r = 0.00$ for all conditions, as indicated in Treatment X (Table 2). This manipulation served to simulate breakdown of a DAD.

Factor B served the function of determining the effect of the nature of the DAD on observer efficiency. At level B_1, subjects were instructed that the DAD was a machine (i.e., a

computer program) that analyzed the given, incomplete information. At level B_2, the subjects were instructed that their DAD was a human co-observer whose responses would be made available to the B_2 subject at the beginning of each trial. B_2 subjects actually saw their co-observers on a monitor throughout the task. All further interaction was minimized.

Subjects at level B_3 were not given any DAD information. They were the ones seen by the B_2 subjects. Unknown to any subjects, the DAD information the B_2 subjects received was not actually generated by the B_3 subjects, but by the computer software, just as it was for the B_1 subjects.

PROCEDURE

The entire experimental procedure was conducted in two sessions, on consecutive days. During each session, a brief rest period was introduced in the middle of the session, after 80 trials. Otherwise, the transition from one trial block to the next was not obvious to the subjects and, in particular, they had no warning that there was a change in the validity of the DAD after the fifth trial block. At the end of the second session, the subjects were debriefed in such a way that the experimental deception was explained to them in complete detail. It was felt that it was important to make this debriefing as extensive as possible so as to elicit the subjects' fullest cooperation in keeping the nature of the experimental task from any potential subjects. During the debriefing, subjects also had an opportunity to convey whether they were aware of the deceptive component of the task. One subject had become aware of this task feature, and this subject's data were excluded from data analysis.

Four subjects were assigned to each of the six conditions involving the use of a DAD, and twelve to the condition in which no DAD was available.

SECTION 3

RESULTS

In their original form, the response measures taken from the subjects consisted of depressions of one of six push buttons on each trial. These data were first transformed into more informative values which were then, in turn, subjected to inferential tests.

SIGNAL DETECTION ANALYSIS

Preliminary data analysis required the transformation of the original frequency data within trial blocks into proportions of "hits" (i.e., occasions on which the subject predicted a collision which actually occurred) and of "false alarms" (i.e., occasions on which the subject predicted a collision when none occurred). In psychophysical terms, the procedure employed in this study is known as a rating procedure. This procedure allowed the subject to make a choice and at the same time indicate how strongly this choice was favored. If only two push buttons had been available, a binary choice could have been made only on one level in any trial. With six push buttons, it was possible to make this binary choice at five levels, thereby yielding five pairs of proportions of hits and false alarms. Each of these five pairs of proportions can be transformed into a single value of d'. This statistic, conceptually, is a measure of the detectability of a signal in an ambient noise distribution. Viewed as a mathematical construct, it is the mean difference between a noise distribution and a signal-plus-noise distribution, in standard deviation units (Green and Swets, 1966, Swets, 1964). While the terms "signal" and "noise", in the earlier developmental stages of signal detection theory, actually were meant to refer to stimulus properties, they are equally applicable to more conceptual levels of processing such as the present task which involves judgment and reasoning. In this task, noise can be considered the equivalent of the supposedly normal random variation in the subject's skills of judgment, the signal strength as the equivalent of task difficulty (or, more precisely, the ease of the task).

The statistic d' was favored for the present data analysis since it is uniquely powerful. It is independent of subject bias, i.e., of the criterion the subject may be using. In fact, the five pairs of proportions make it possible to determine five values of d', at five different criteria. Except for error variation, all five values of d' can be assumed to be the same. In the actual data analysis, only a single value of d', instead of five, was computed, by a method involving a simple linear regression procedure (Green and Swets, 1966). All subsequent analyses were performed with values of d' as dependent variables.

Four treatment combinations operated on the subjects. These combinations are defined by trial outcome (collision or safety) and DAD prediction (again collision or safety). In coded form, these treatment combinations are presented in Table 3. Using these codes for reference, it can be observed that the overall decision making performance of the subjects is a function of the proportion of collisions that subjects

TABLE 3. Codes for the Four Possible Trial Conditions

Prediction of DAD	Trial Outcome: Collision	Trial Outcome: Safety
Collision	A	B
Safety	C	D

predicted in cells A and C, compared to the proportion of collisions they predicted in cells B and D. The measure of sensitivity that is based on this contrast is called d'(sen) in the present context, and it can be understood as

$$d'(\text{sen}) = d'_{(A,C)-(B,D)}.$$

This measure of performance is only one of a large number of possible measures. The reliance of a subject on the DAD is based on the contrast between proportions of

collision predictions in cells A and B, compared to cells C and D. The corresponding measure of reliance, d'(rel), is

$$d'(\text{rel}) = d'_{(A,B)-(C,D)}.$$

A highly efficient subject would be one who agreed with the DAD whenever it was right, and disagreed with it on those occasions when it was wrong. In other words, this subject would obtain a high positive value for d'_{A-D} (which would reflect *concurrence* with a correct DAD) and a high negative value for d'_{B-C} (which would be positive for subjects who would blindly follow the predictions of an incorrect DAD). A measure of subject efficiency, d'(eff), can therefore be obtained as the difference between these two values:

$$d'(\text{eff}) = d'_{A-D} - d'_{B-C}.$$

These three measures, d'(sen), d'(rel), and d'(eff), were used in the subsequent data analysis.

PERFORMANCE AS A FUNCTION OF DAD CHARACTERISTICS

Statistical analyses of all major effects were conducted at several levels of detail, and the least detailed ones that would preserve all relevant information were chosen for this report. Regarding the independent variable of trial blocks, the usual phenomenon occurred, namely that there was a substantial learning effect over the first five blocks. This effect did not interact in a significant manner with any other variable. Since this effect in itself is of no interest in the context of this report, it was regarded more informative to consider the first two trial blocks as a warm-up and practice session, and to view blocks 3 to 5 as occasions of practiced performance with a valid DAD (or, in the case of the control group, practiced performance without a DAD) and blocks 6 to 8 as performance with a DAD that had lost its validity. This approach also permits a simpler statistical analysis.

Performance efficiency of subjects with a valid DAD in trial blocks 3 to 5 was significantly superior to that of control subjects without a DAD ($F=8.40$, $df=1,29$, $p<.05$), and it was reduced significantly below that of control subjects when the DAD became invalid in blocks 6 to 8 ($F=6.44$, $df=1,29$, $p<.05$). Figure 2 presents the underlying values of d'(sen). The continuous curves in this figure and subsequent ones represent least squares fitted growth functions.

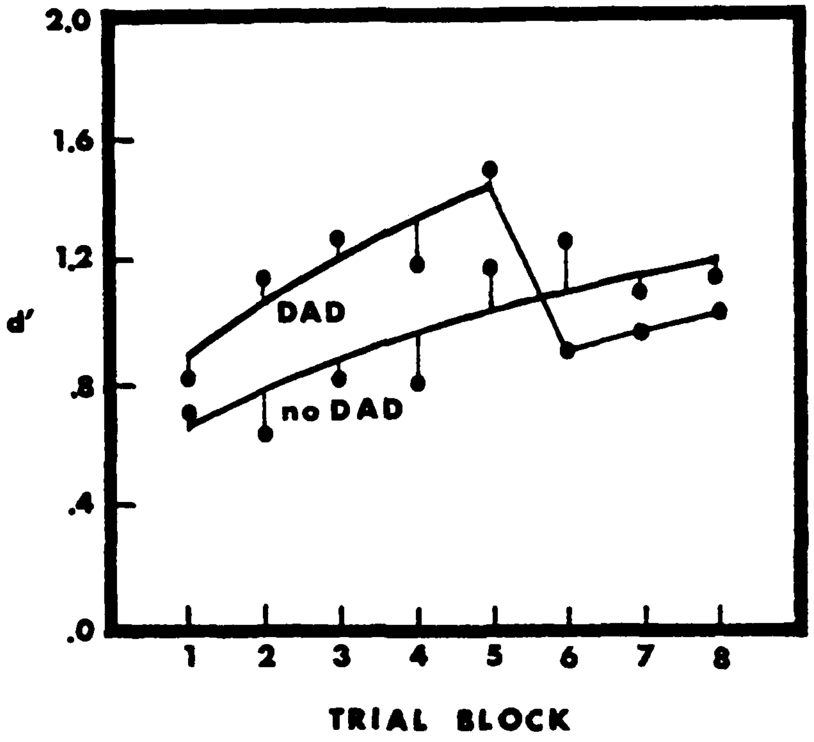


FIGURE 2. Performance as measured in d'(sen), as a function of trial blocks, with and without DAD.

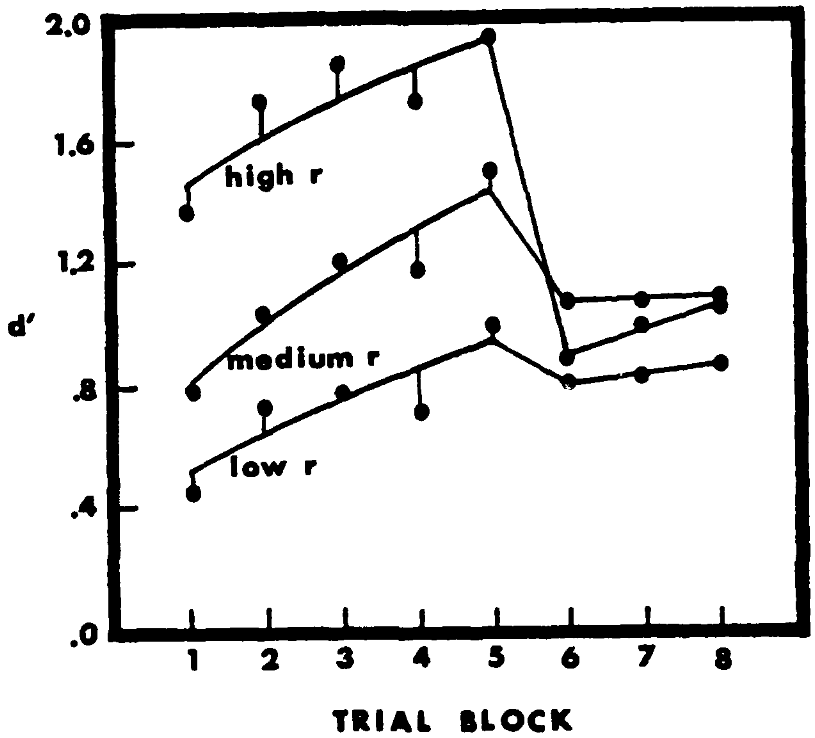


FIGURE 3. Performance as measured in d'(sen), over trial blocks, as a function of DAD validity (r).

Figure 3 shows performance over trial blocks, in terms of d'(sen), as a function of DAD reliability. The treatment effect of validity is significant in blocks 3 to 5 ($F=9.85$, $df=2,18$, $p<.05$), and so is the overall drop in performance once the DAD is invalid ($F=13.72$, $df=1,18$, $p<.05$). The more validity the DAD initially had, the more severe was the drop in d'(sen), resulting in the interaction between initial validity and blocks 3 to 5 versus 6 to 8 ($F=8.10$, $df=2,18$, $p<.05$). Previous DAD validity did not have a significant effect on subsequent performance with an invalid DAD.

Further review of Figure 3 reveals that subjects showed their highest sensitivity to the task on trial block 5, the last block in which the DAD was still valid. Performance at this point, while not asymptotic, showed the subjects at the best they achieved. This performance has to be understood in the context of a baseline. One such baseline is provided by the control subjects (cf. Figure 2), but another one can be determined theoretically. The predictive validity of a DAD can be expressed as a value of d'. Table 4 contrasts these theoretical values of d' with those achieved by the subjects

TABLE 4. Observer Performance, d'(sen), on Trial Block 5, Compared with the Theoretical Sensitivity of the DAD, d'(DAD)

DAD validity, (r)	d'(DAD)	d'(sen), block 5
.40 (low)	1.05	.99
.60 (medium)	1.68	1.51
.80 (high)	2.56	1.95
.00 (control)	.00	1.17

on trial block 5. Two aspects of this table are noteworthy. First, observers with a DAD of low validity ($r = .40$) performed more poorly ($d' = .99$) than control subjects who did

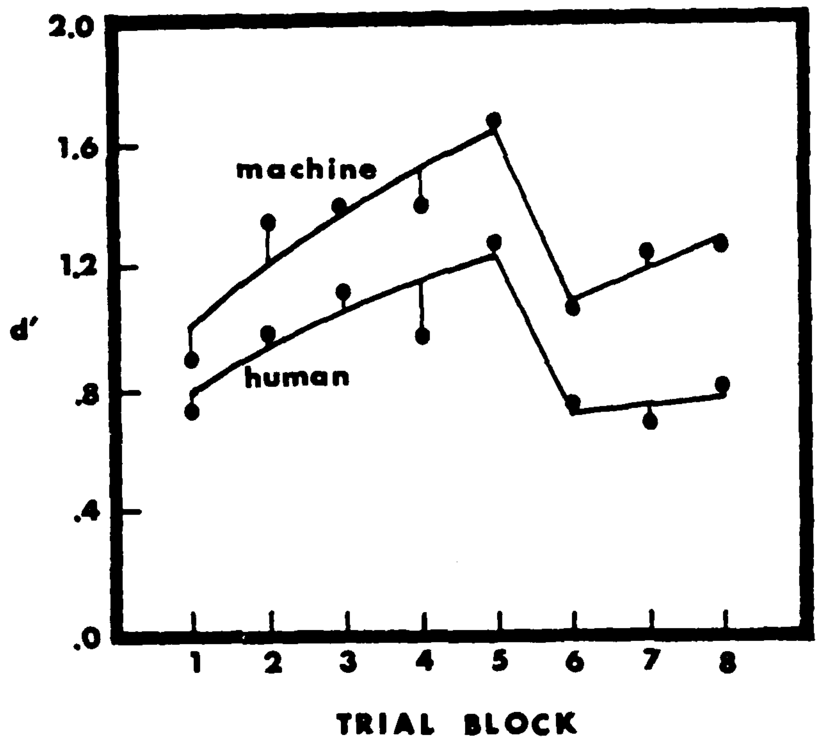


FIGURE 4. Performance as measured in d'(sen), over trial blocks, as a function of DAD type.

not have a DAD at all (d' = 1.17). Second, mean performance of subjects under all three conditions employing a DAD was inferior to the theoretical sensitivity values of the DAD. These theoretical values would have been achieved by a subject who would have followed the DAD prediction blindly on all trials.

Figure 4 summarizes the same performance over trial blocks, again in terms of d'(sen), but as a function of DAD type. Despite the non-overlap of these functions, DAD type had no significant effect, nor did it interact significantly with any other variables in regard to d'(sen). In particular, there was no interaction between DAD type and DAD validity.

RELIANCE ON THE DAD

It is of interest to what degree the performance of subjects might be understood in terms of their reliance on the DAD. Obviously, reliance on a valid DAD produces high performance even if the subject does nothing more than to agree with the DAD. To determine the patterns of reliance on the DAD, the values of d'(rel) were analyzed, as a function of DAD validity and of DAD type, and again over trial blocks 3 to 5 and 6 to 8, as previously. These analyses showed that reliance on the DAD, blocks 3 to 5, was a function of DAD validity ($F=13.70$, $df=2,18$, $p<.05$), and that it was not significantly affected by DAD type. As was to be expected, there was a considerable drop in reliance under all treatment conditions after trial block 5, when the DAD became invalid ($F=194.80$, $df=1,18$, $p<.05$). This drop was characterized by a significant interaction: The decrease in reliance on the DAD, once it failed, was most pronounced if the DAD had been a high-validity machine or a low-validity human; it was least pronounced if the DAD had been a low-validity machine or a high-validity human ($F=4.31$, $df=2,18$, $p<.05$). This interaction was an orderly function of the degree of DAD validity, as a linear trend component of the interaction suggests ($F=5.16$, $df=1,18$, $p<.05$).

Covariance analyses of d'(sen), with d'(rel) as a covariate, yielded no further insights. The only difference between these results and those obtained in the analysis of variance of d'(sen) was that the covariance analysis did not yield a significant effect of DAD validity, as was to be expected.

SUBJECT EFFICIENCY

Subject efficiency, calculated by the measure d'(eff), required a different method of analysis than d'(sen) and d'(rel). The measure d'(eff) is identical with d'(sen) when DAD validity is zero. The higher DAD validity is, the more d'(eff) takes on an independent meaning. For this reason, d'(eff) was analyzed only for blocks 3 to 5. An analysis of variance yielded no significant effect at all, neither for DAD type nor for DAD validity, nor for the interaction. This outcome is inconclusive, of course, but it is interesting in that it suggests that efficiency is indeed independent of the treatment condition, as it should be. In other words, subject efficiency may be an individual difference variable that remains stable regardless of treatment conditions.

This finding was considered most important in the present context since individual differences between subjects were indeed quite pronounced. After all, the apparent difference in performance as a function of DAD type (cf. Figure 4) was non-significant due to large between-subject variances. Also the central question what factors might predict performance once the DAD had lost its validity was still unresolved. The independent variables of the study, after all, did not have a demonstrable effect on performance in blocks 6 to 8.

To determine whether other variables might be operating on subjects' behavior, a correlational analysis was conducted, using all subjects except the control group, and employing within-treatment correlation coefficients. Three particular coefficients are of interest, namely those between d'(eff) for blocks 3-5, d'(sen) for blocks 3-5, and d'(sen) for blocks 6-8. The last one of these variables is the one for which a predictor variable would be of primary interest. The correlation matrix is presented in Table 5a. All three of the correlation coefficients are quite high, and one might conclude that both d'(sen) for blocks 3-5 and d'(eff) for blocks 3-5 are good predictors of d'(sen) for blocks 6-8. A partial correlation analysis, however, suggests a different interpretation. As shown in Table 5b, only the correlation between d'(eff) for blocks 3-5 and d'(sen) for blocks 6-8 is substantial.

TABLE 5. (a) Correlation Coefficients and (b) Partial Correlation Coefficients between Three Behavioral Variables

(a)

	d'(eff),3-5	d'(sen),3-5
d'(sen),3-5	.835	
d'(sen),6-8	.945	.838

(b)

	d'(eff),3-5	d'(sen),3-5
d'(sen),3-5	.240	
d'(sen),6-8	.816	.275

SECTION 4

DISCUSSION

Any study that attempts to simulate a complex environmental situation is bound to encounter serious restrictions in the degree to which its results can be generalized. The current study is no exception. There can be little question that any generalization from the results of this study is limited by the degree of comparability between parameters of this study and conditions in the setting to which the inference is to be applied. It is particularly noteworthy that the task that subjects had to carry out, while it did involve judgmental operations, is but remotely similar to the taxing demands of a C^3-situation. What could not be simulated in the context of an experiment was a sense of urgency that judgment errors be avoided at all cost. Similarly, the decision aids used in this study were more contrived than they are in a concrete C^3-task. With these restrictions in mind, however, several cautious conclusions may be drawn.

The study suggests that the performance of C^3-systems with a valid decision aid is, on the average, superior to the performance of comparable systems without the benefits of a decision aid. However, this statement applies only as long as the validity of the decision aid is relatively high. If the validity is low, C^3-system performance is inferior to the performance that would be achieved without the use of a decision aid. This finding is intuitively agreeable; if a decision aid has sufficiently low validity, it comes to the point of being a hindrance rather than a help, and the user derives negative value from its use.

The first phase of training offered to the subjects (trial blocks 1 to 5) was not long enough to achieve anything like asymptotic behavior. Nonetheless, at this point the subjects were well into their second session of experience with the task, and it seemed unlikely that a substantial improvement in their performance would be forthcoming. The terminal performance of subjects, with the support of a valid DAD, was substantially inferior to the performance they could have achieved if they had followed a maximizing strategy, following the DAD consistently and blindly. This finding is both sobering and consistent with the literature. There is *abundant* evidence that subjects fail to follow a maximizing strategy after extensive training, even if they are fully aware of the contingencies

and of the fact that they could do better if they maximized rather than engaged in some form of gamble (Edwards, 1956, 1961, Estes, 1964, Goodnow and Postman, 1955, Hoffeld and Webster, 1964, Kahneman, Slovic, and Tversky, 1982, Rubinstein, 1959). The only comfort that can be derived from this very persistent human pattern is that it pays off when the DAD breaks down. At this point, an observer who would have maximized would be useless to the system. Even if the observer would be capable of learning that independent judgment was suddenly required, the skills needed to do so now would have to be developed for the first time.

When a previously valid DAD breaks down, the performance of a C^3-system is degraded to a level lower than that of persons who have been operating without a DAD throughout. This drop in performance reaches the same low level regardless of the prior validity of the DAD. In other words, the validity of a DAD does not affect performance of the human operator once the DAD becomes invalid. This finding holds for DADs in the form of human co-observer information and of machine information alike.

None of the treatment combinations of independent variables that were employed in this study seemed to affect operator performance in a differential way once the DAD had become invalid. At the same time, there was evidence that operators differed considerably in the quality of their performance once they were operating with an invalid DAD. It was recognized that it would be of considerable theoretical interest and practical relevance to find a predictor variable for performance with an invalid DAD. Since this performance was not affected by the independent variables, it was assumed to be a subject variable, i.e., a variable that is relatively stable within individuals.

As indicated before, neither performance with a valid DAD nor reliance on a DAD turned out to be informative predictors of performance with an invalid DAD. Subject efficiency, operationalized as d'(eff), on the other hand, did turn out as a strong predictor. This variable, as suggested previously, reflects a subject's ability to concur with a DAD on those occasions when it is correct and at the same time to disagree with it when it is incorrect. Such an ability, in more general terms, makes for success in any enterprise, and it is not surprising that it should work in this instance, too.

The suggestion which emerges from this analysis, of course, is that the establishment of a C^3-system of high quality is preeminently a matter of selection or possibly training. Lest this suggestion be adopted too enthusiasti-

cally, it must be pointed out that this construct requires independent validation. At present, subject efficiency has been shown merely to be correlated with the variable of interest; it has not been established as a sufficient or even a necessary antecedent. A first step in this direction was undertaken in the present study. AFOQT scores, including subtest scores, were obtained for all subjects wherever possible. The yield was insufficient for a correlational analysis, but such an analysis, if any, might be a useful direction for future work.

The work undertaken here was not intended to lead in the direction which was eventually indicated. For this reason, the suggestion to examine efficiency as a potential individual difference variable should be taken as intended, namely as a recommendation for further study. More efforts also seem appropriate to validate the findings of this study in operational settings where possible.

SECTION 5

REFERENCES

Asch, S.E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological Monographs, 70, 9, No. 416.

Bateson, N. (1966). Familiarization, group discussion, and risk taking. Journal of Experimental Social Psychology, 2, 119-129.

Edwards, W. (1956). Reward probability, amount, and information as determiners of sequential two-alternative desisions. Journal of Experimental Psychology, 52, 177-188.

Edwards, W. (1961). Probability learning in 1000 trials. Journal of Experimental Psychology, 62, 385-394.

Estes, W.K. (1964). Probability learning. In A. Melton (Ed.), Categories of human learning. New York: Academic Press.

Festinger, L. Riecken, H.W., and Schachter, S. (1956). When prophesy fails. New York: Harper and Row.

Fischer, G.W., Edwards, W., and Kelly, C.W. (1978). Decision theoretic aids for inference, evaluation, and decision making: A review of research and experience. Technical Report TR-78-1-30, AFAMRL.

Freedy, A., Davis, K.B., Steeb, R., Samet, M.G., and Gardiner, P.C. (1976). Adaptive computer aiding in dynamic decision processes: Methodology, evaluation, and applications. Technical Report PFTR-1016-76-8/30. Woodland Hills, Calif.: Perceptronics.

Goodnow, J.J., and Postman, L. (1955). Probability learning in a problem solving situation. Journal of Experimental Psychology, 49, 16-22.

Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Greenberg, J. (1979). Group versus individual equity judgments: Is there a polarization effect? Journal of Experimental Social Psychology, 15, 504-512.

Hill, G.W. (1982). Group versus individual performance: Are N+1 heads better than one? Psychological Bulletin, 91, 517-539.

Hoffeld, D.R., and Webster, R.L. (1964). Effect of altered likelihood values upon information use. Psychological Reports, 15, 79-82.

Kahneman, D., Slovic, P., and Tversky, A., Eds. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

Maier, N.R.F. (1970). Problem solving and creativity in individuals and groups. Belmont, Calif.: Brooks-Cole.

Pruitt, D.G. (1971). Choice shifts in group discussion: An introductory review. Journal of Personality and Social Psychology, 20, 339-358.

Rubinstein, E. (1959). Some factors in probability matching. Journal of Experimental Psychology, 57, 413-416.

Shiflett, S. (1979). Toward a general model of small group productivity. Psychological Bulletin, 86, 67-79.

Swets, J.A., Ed. (1964). Signal detection and recognition by human observers: Contemporary readings. New York: Wiley.

Wallach, M.A., and Kogan, N. (1965). The roles of information, discussion, and consensus in group risk taking. Journal of Experimental Social Psychology, 1, 1-19.

Winer, B.J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

U.S. GOVERNMENT PRINTING OFFICE 1987

END
Feb.
1988
DTIC